Паксос в картинках

Константин Осипов
kostja@scylladb.com





Обо мне

- ex-MySQL core dev
- ex-CTO Tarantool
- Автор курсов по Техносфера@ВМК
- Multi-Paxos и Raft в ScyllaDB



Краткая история Paxos

1978 - Time, Clocks and the Ordering of Events in a Distributed System

1982 - The Byzantine Generals Problem - 1987 - The Byzantine Generals

1990 - Part time Parliament

1999 - Paxos Made Simple

2011 - Paxos made moderately complex

2020 - Paxos vs Raft: Have we reached consensus on distributed consensus?



Где используется Paxos

Paxos	**Raft**
Zookeeper, Cassandra, Aerospike, MySQL (**)	Consul, etcd, CockroachDB, YugaByte, TiDB, MongoDB(*), Tarantool (*), RavenDB

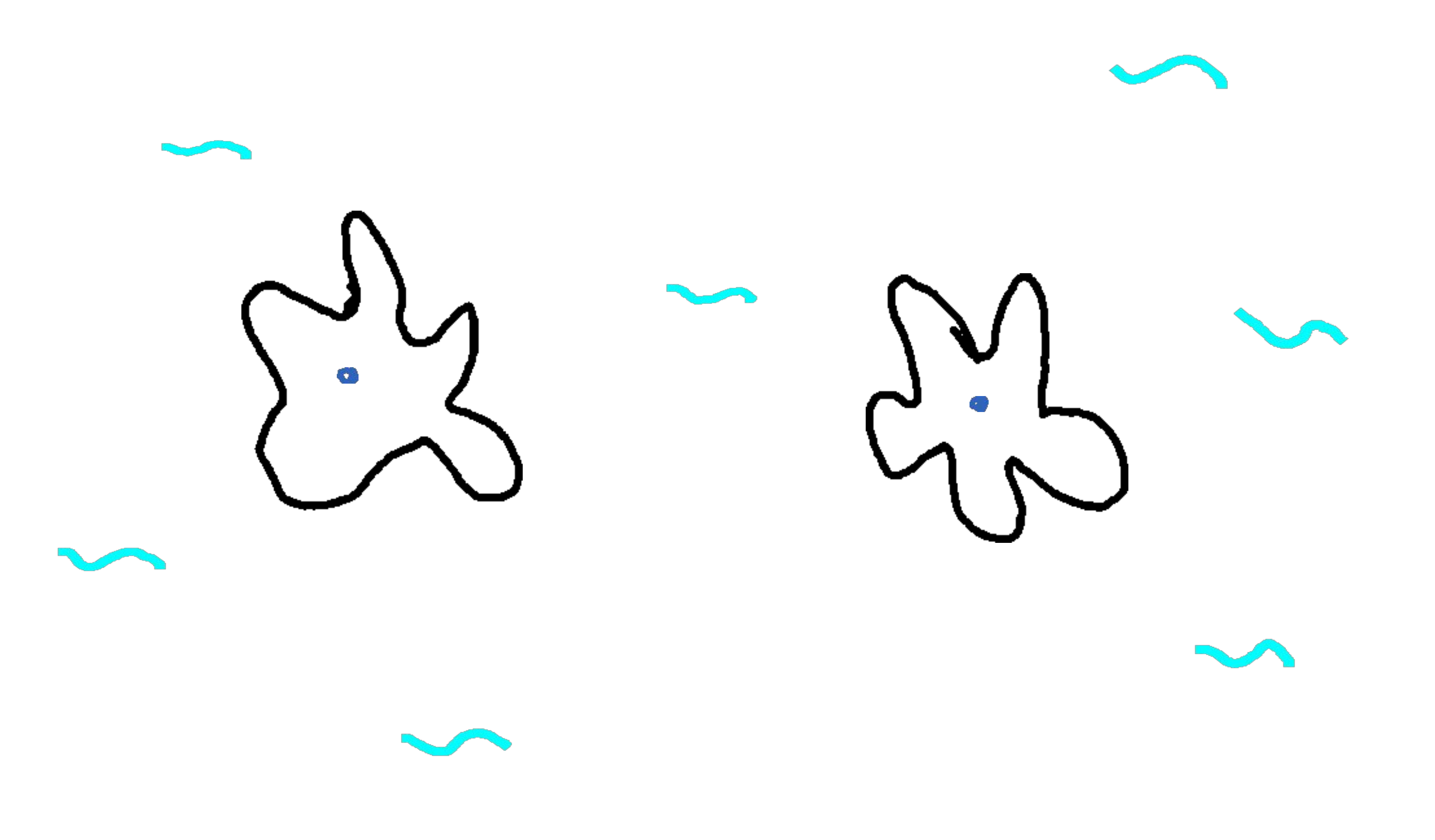
(*) - используется свой велосипед, названный Raft
(**) - MySQL Group Replication

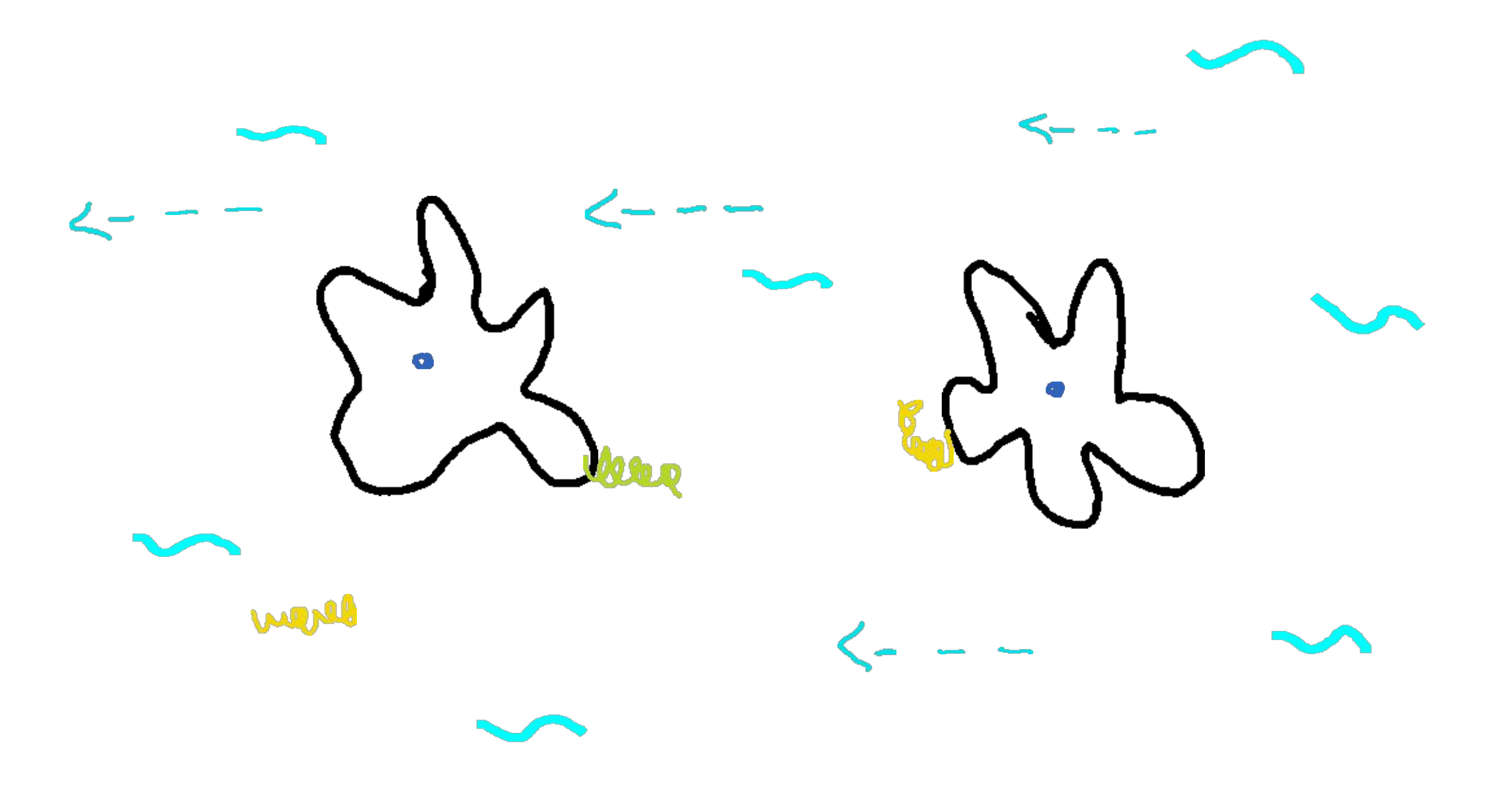


Задача

- Only a value that has been proposed may be chosen,
- Only a single value is chosen, and
- A process never learns that a value has been chosen unless it actually has been.











In the absence of failure or message loss, we want a value to be chosen even if only one value is proposed by a single proposer. This suggests the requirement:

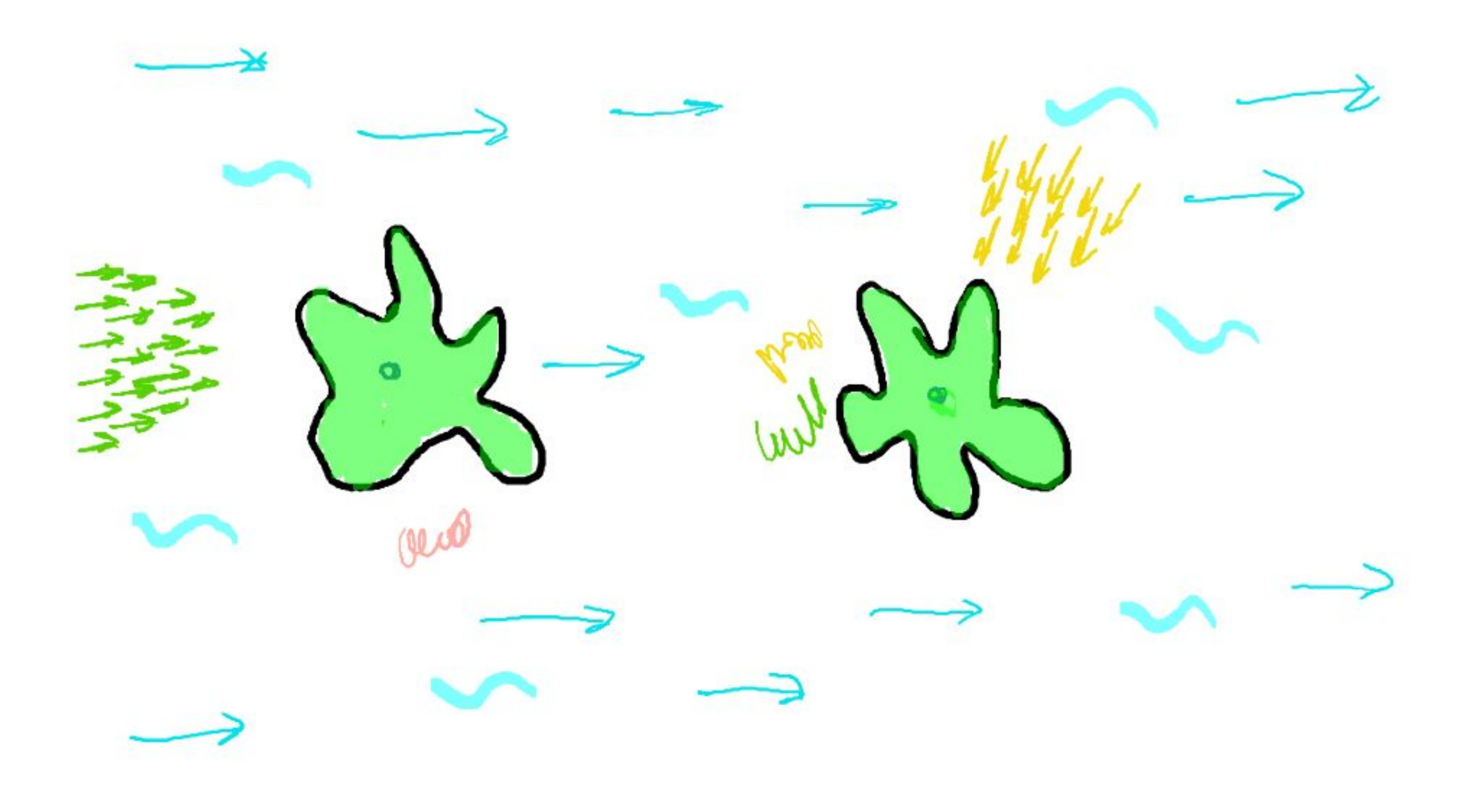
P1. An acceptor must accept the first proposal that it receives.

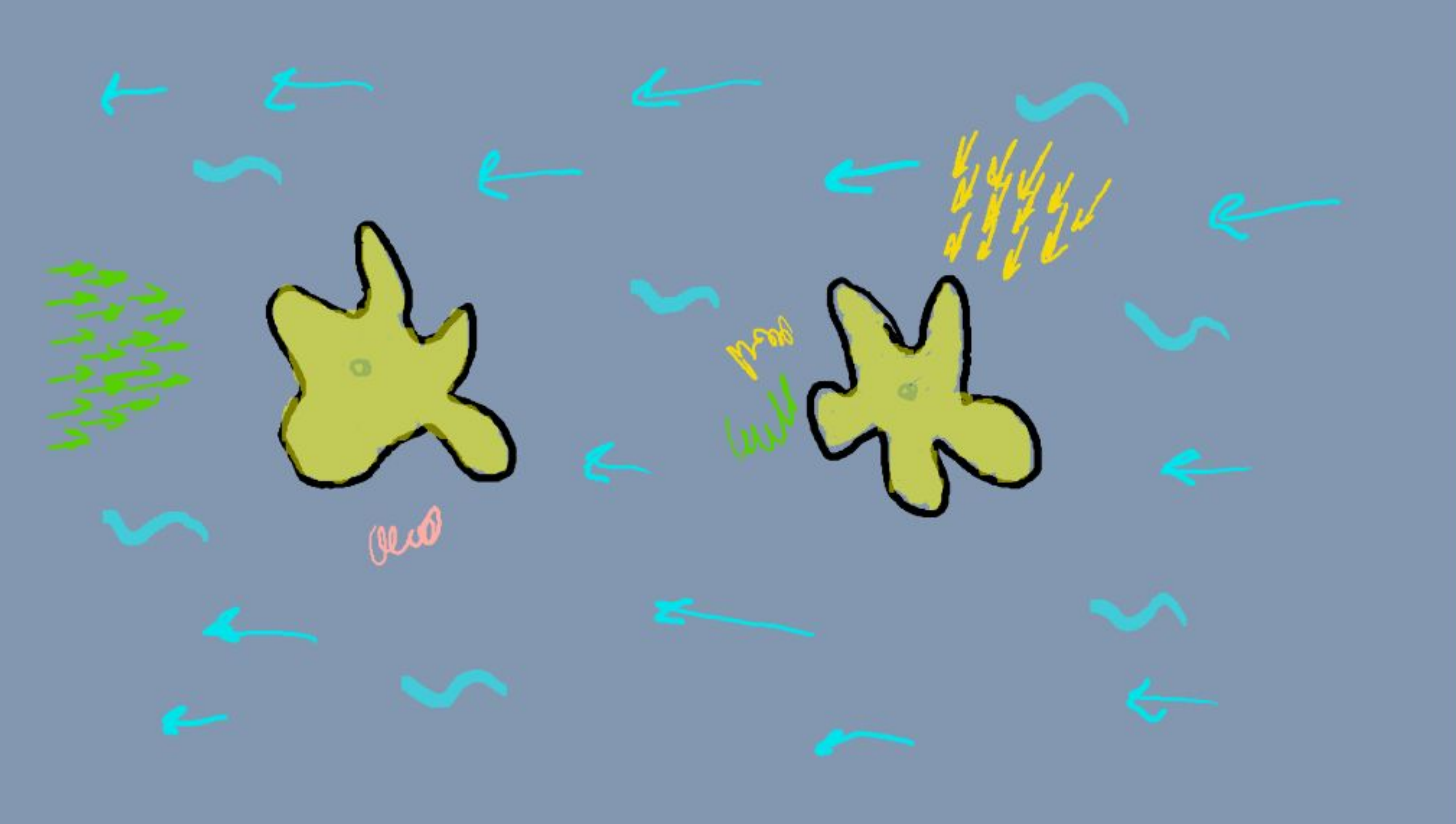


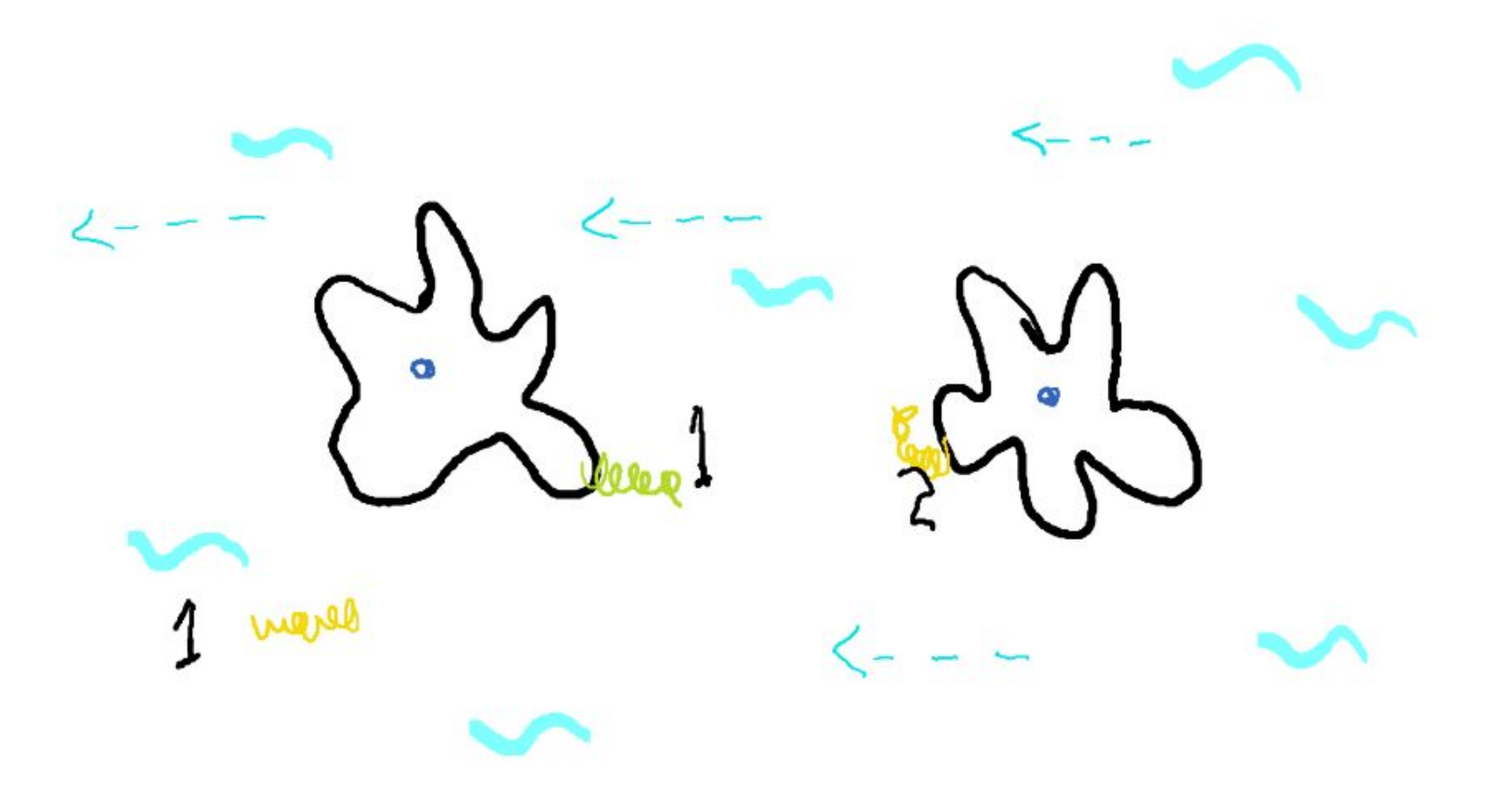


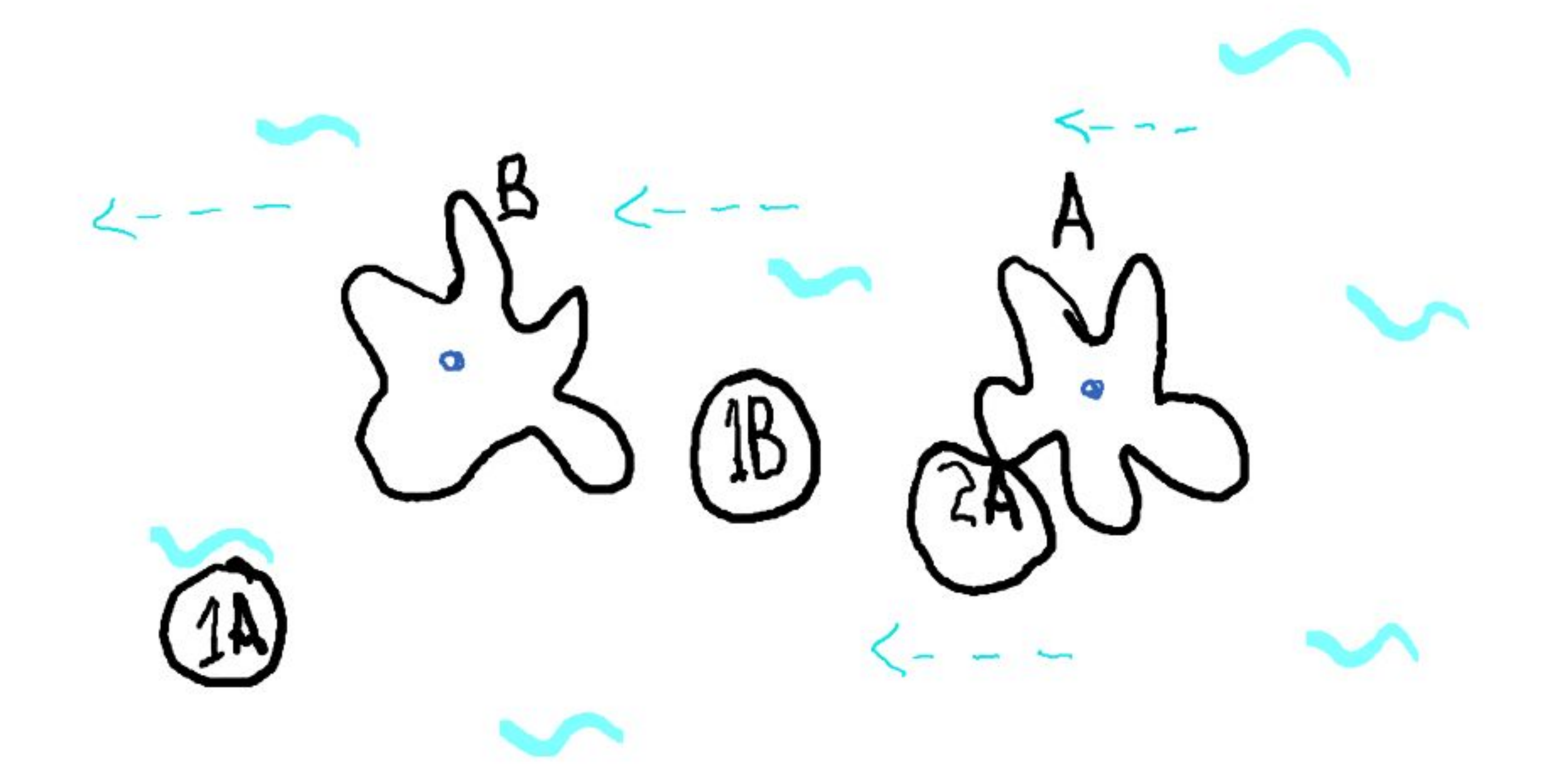


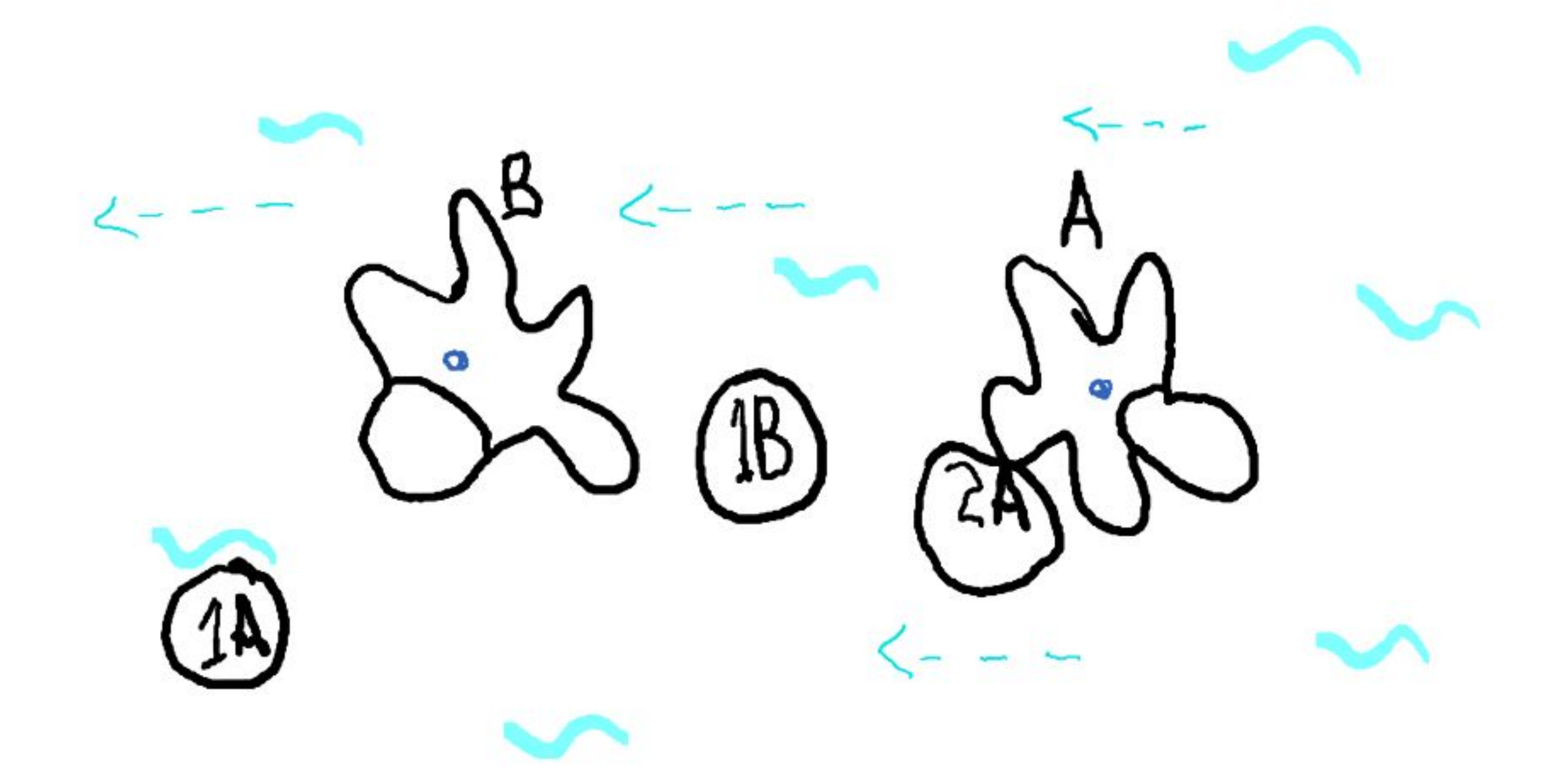









B
A
1B
2A
1A

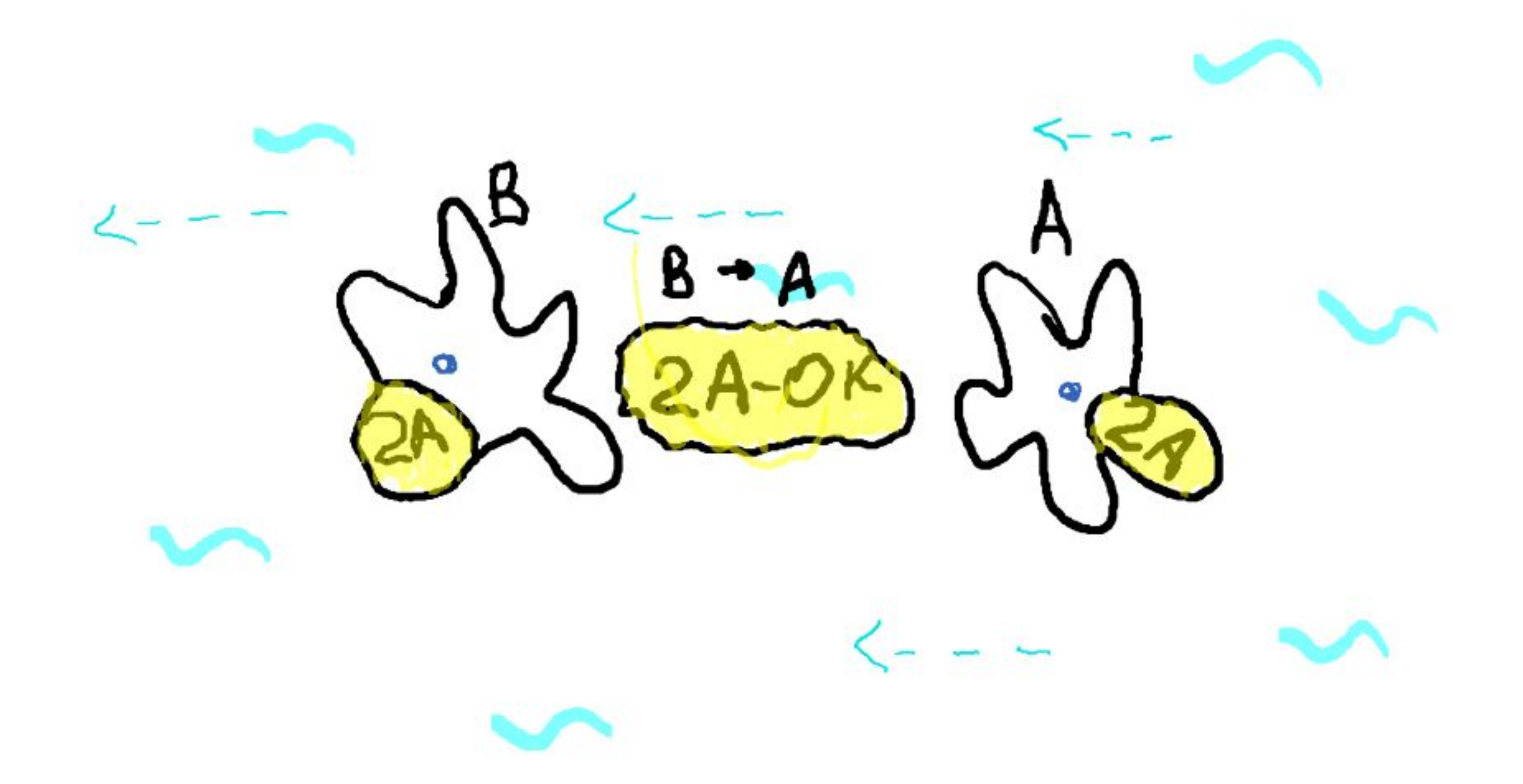

B
A
1B
2A
1A

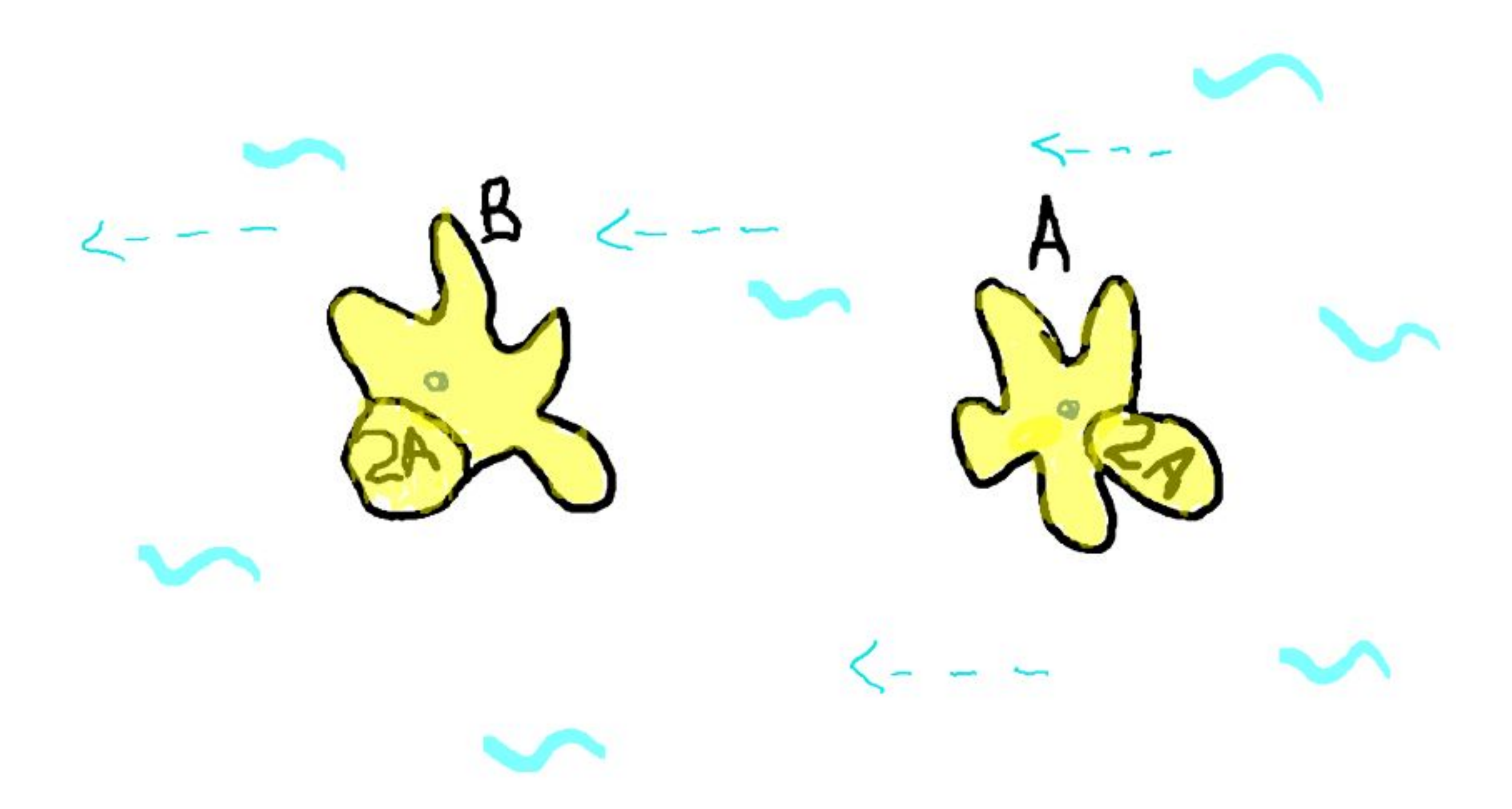

B
A
1B
1B
2A
2A
1A

1B<2A

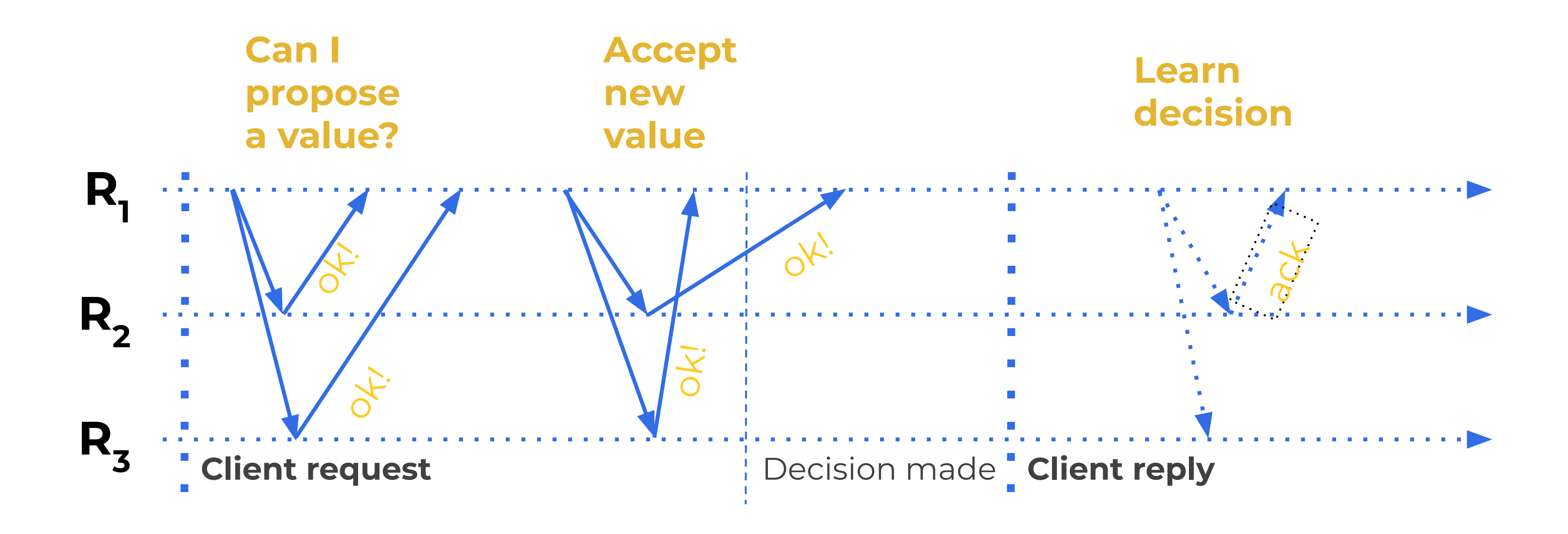



B
B ← A
A
2A
2A
2A


B
B → A
2A-OK
A
2A
2A


B
2A
A
2A

Что может пойти не так



TLA+

```
int i ;
void main()
  { i = someNumber() ;
    i = i + 1 ;
  }
```

$$
\begin{aligned}
\lor\ & \land\ pc = \text{“}start\text{”} \\
      & \land\ i' \in 0\,..\,1000 \\
      & \land\ pc' = \text{“}middle\text{”} \\
\lor\ & \land\ pc = \text{“}middle\text{”} \\
      & \land\ i' = i + 1 \\
      & \land\ pc' = \text{“}done\text{”}
\end{aligned}
$$

We need nondeterminism to describe systems,

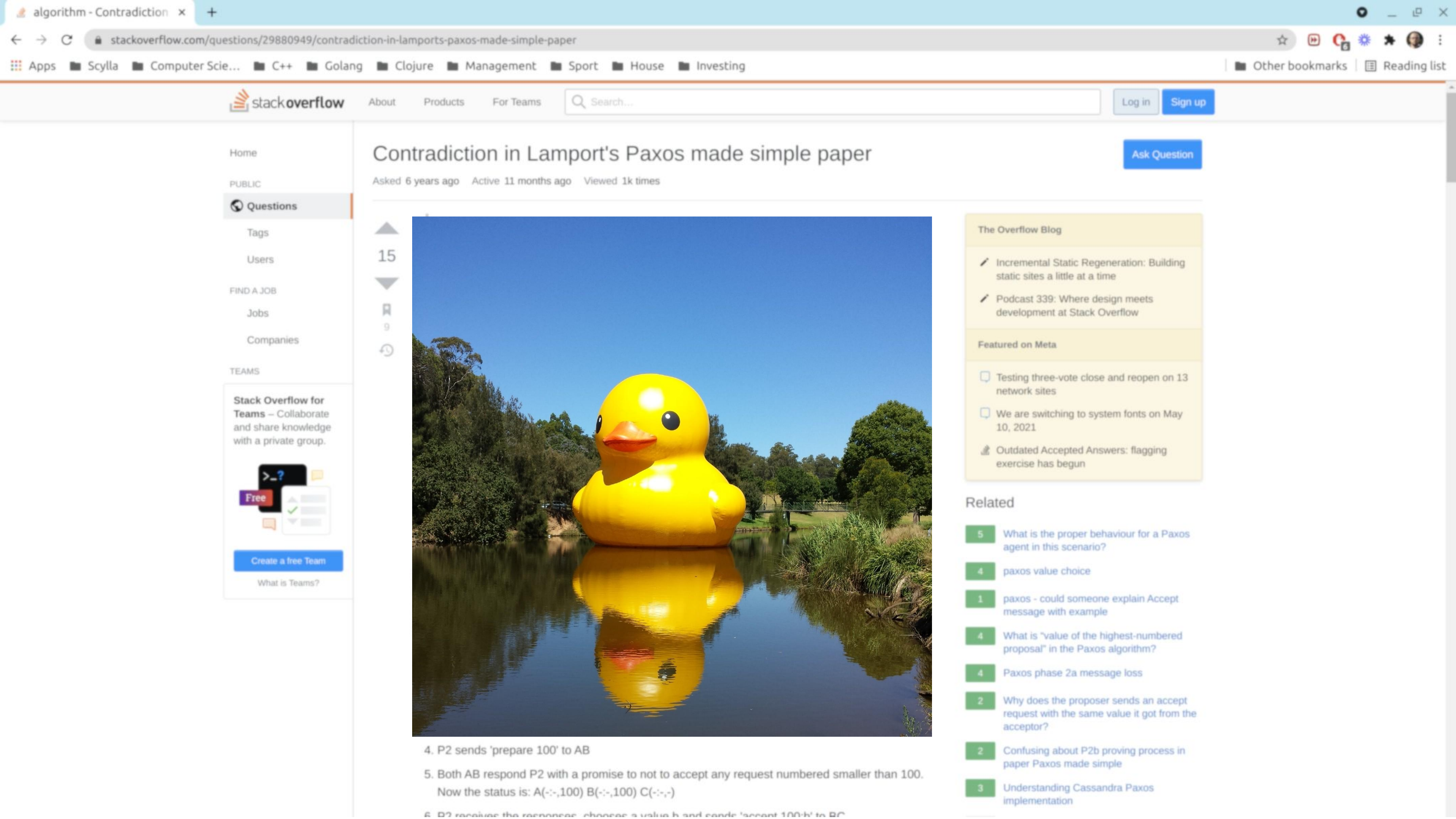
because we can't predict in what order things happen.

Leslie Lamport quote



A few years ago an amazon engineer (a smart guy so when he said something I believed it) told me that there is a sentence in “Paxos Made Simple” which could be misinterpreted and there are three implementations of Paxos (he knows them) that contain that error.




algorithm - Contradiction
stackoverflow.com/questions/29880949/contradiction-in-lamports-paxos-made-simple-paper
Apps
Scylla
Computer Scie...
C++
Golang
Clojure
Management
Sport
House
Investing
Other bookmarks
Reading list
stackoverflow
About
Products
For Teams
Search...
Log in
Sign up
Home
PUBLIC
Questions
Tags
Users
FIND A JOB
Jobs
Companies
TEAMS
Stack Overflow for Teams – Collaborate and share knowledge with a private group.
Free
Create a free Team
What is Teams?
Contradiction in Lamport's Paxos made simple paper
Ask Question
Asked 6 years ago Active 11 months ago Viewed 1k times
15
9
4. P2 sends 'prepare 100' to AB
5. Both AB respond P2 with a promise to not to accept any request numbered smaller than 100. Now the status is: A(-:-,100) B(-:-,100) C(-:-,-)
The Overflow Blog
Incremental Static Regeneration: Building static sites a little at a time
Podcast 339: Where design meets development at Stack Overflow
Featured on Meta
Testing three-vote close and reopen on 13 network sites
We are switching to system fonts on May 10, 2021
Outdated Accepted Answers: flagging exercise has begun
Related
5 What is the proper behaviour for a Paxos agent in this scenario?
4 paxos value choice
1 paxos - could someone explain Accept message with example
4 What is "value of the highest-numbered proposal" in the Paxos algorithm?
4 Paxos phase 2a message loss
2 Why does the proposer sends an accept request with the same value it got from the acceptor?
2 Confusing about P2b proving process in paper Paxos made simple
3 Understanding Cassandra Paxos implementation

Что осталось за скобками

- Удаление состояния
- Мульти-паксос
- Идентификация участников кластера
- Начальный состав участников
- Изменение состава участников
- Выход из строя клиента
- E-Paxos, FastPaxos, CasPaxos, Flexible Paxos, Paxos Quorum Leases
- ...



Спасибо!

